# PREDICTING THE BILLBOARD TOP 100: APPLYING MACHINE LEARNING TO MUSIC STREAMING DATA

**DSC288R Group 2**

Lindsay Sager
Tiffany Tong

# AGENDA

- Background
- Literature Survey
- Why Machine Learning
- Dataset Pipeline
- Feature Extraction
- Details on Models Used
- Results and Observations
- Next Steps and Risks
- References

# BACKGROUND

## Objective of Our Research:

- Predict whether a song has reached the Billboard Top 100 based exclusively on its audio features

## Value to the Music Industry:

- Tailoring music to audience preferences
- Investment in music rights
- Understanding trends

# LITERATURE SURVEY

## Related research fields:

- Music Information Retrieval (MIR)
- Hit Song Science (HSS)

## What people tried to solve this problem:

- Logistic Regression model
- Decision trees
- Random Forest
- K-Nearest Neighbor
- Support vector machines
- Neural networks

# LITERATURE SURVEY (CONT.)

Takeaways from Literature Surveys:

- Importance of data quality
- Models capable of capturing non-linear patterns (i.e. tree-based, neural nets) are more successful

Gaps to address:

- Time-series trends in hit song audio feature composition
- Analysis of feature importance in determining hit song prediction

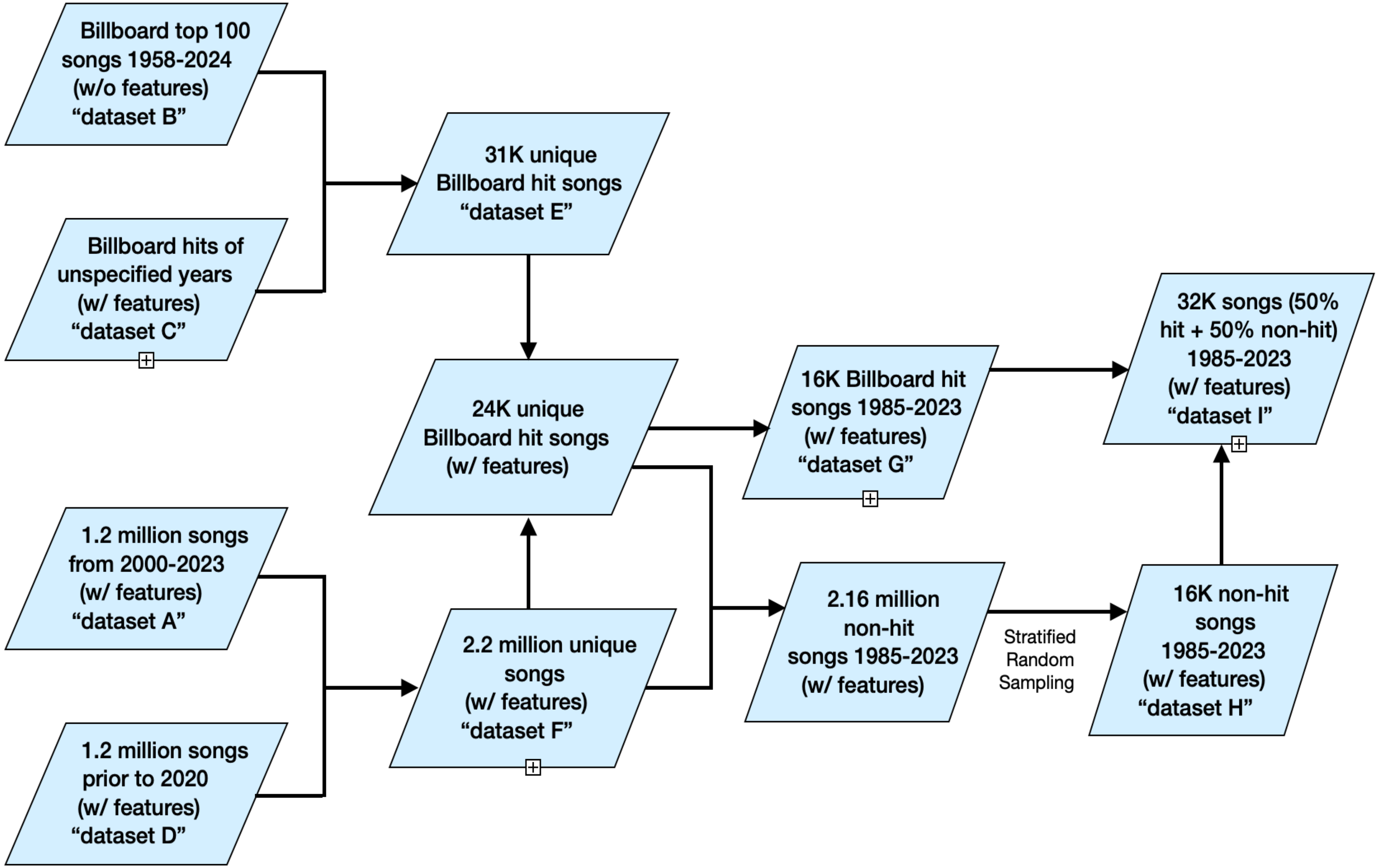
# WHY MACHINE LEARNING

- Machine Learning (ML) algorithms can identify complex patterns in large datasets.
- Song features can be numerous, subtle, and even abstract, such as "danceability" or "energy".

# DATASET PIPELINE

Source Data:

1. Source of truth for Billboard Top 100
2. Billboard data w/ audio features
3. 1.2M songs w/ audio features
4. 1M songs w/ audio features



# DATASET OVERVIEW (CONT.)

16,242 hit songs w/ features

(1985-2023)

+

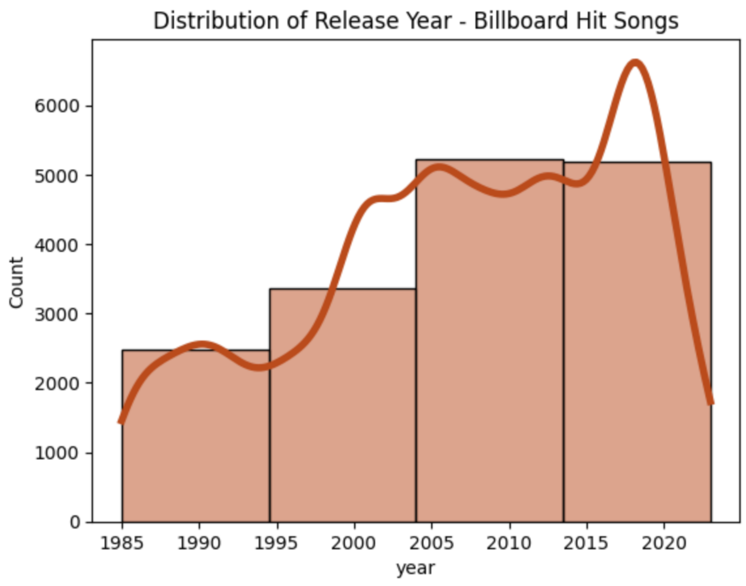
16,242 non-hit songs w/ features

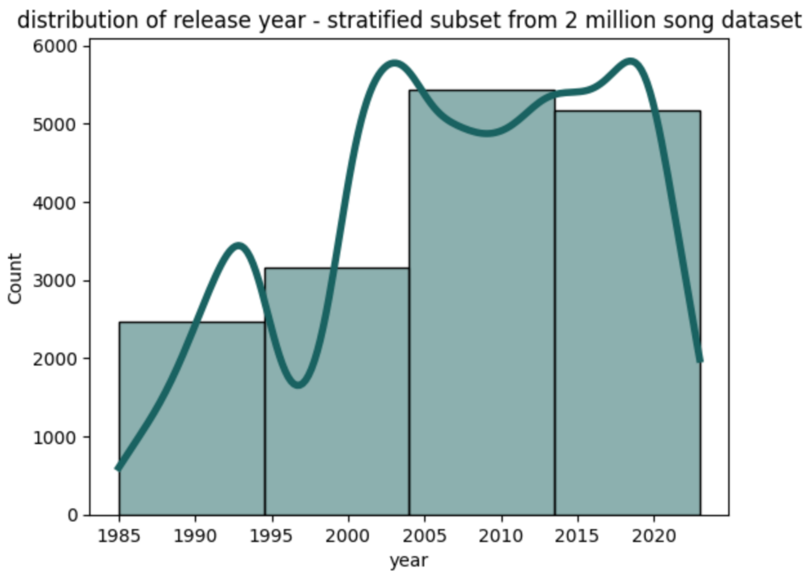
(1985-2023)

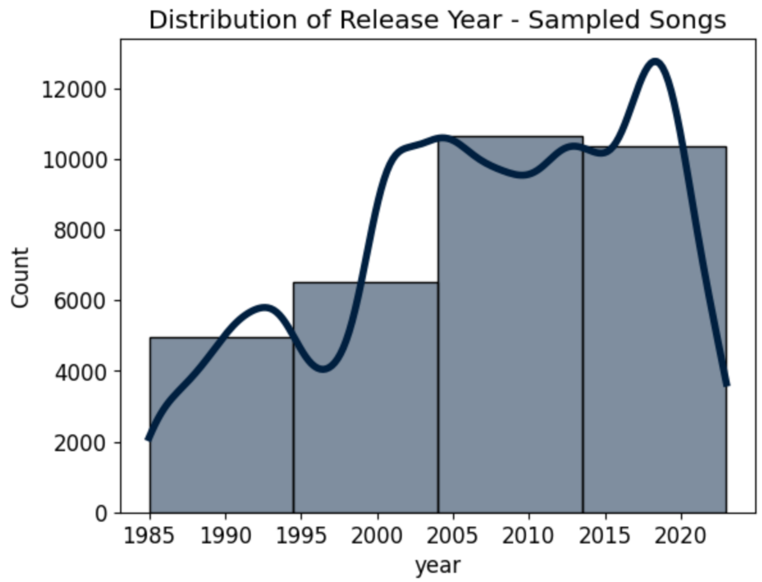
=

32,484 songs w/ features

(1985-2023)



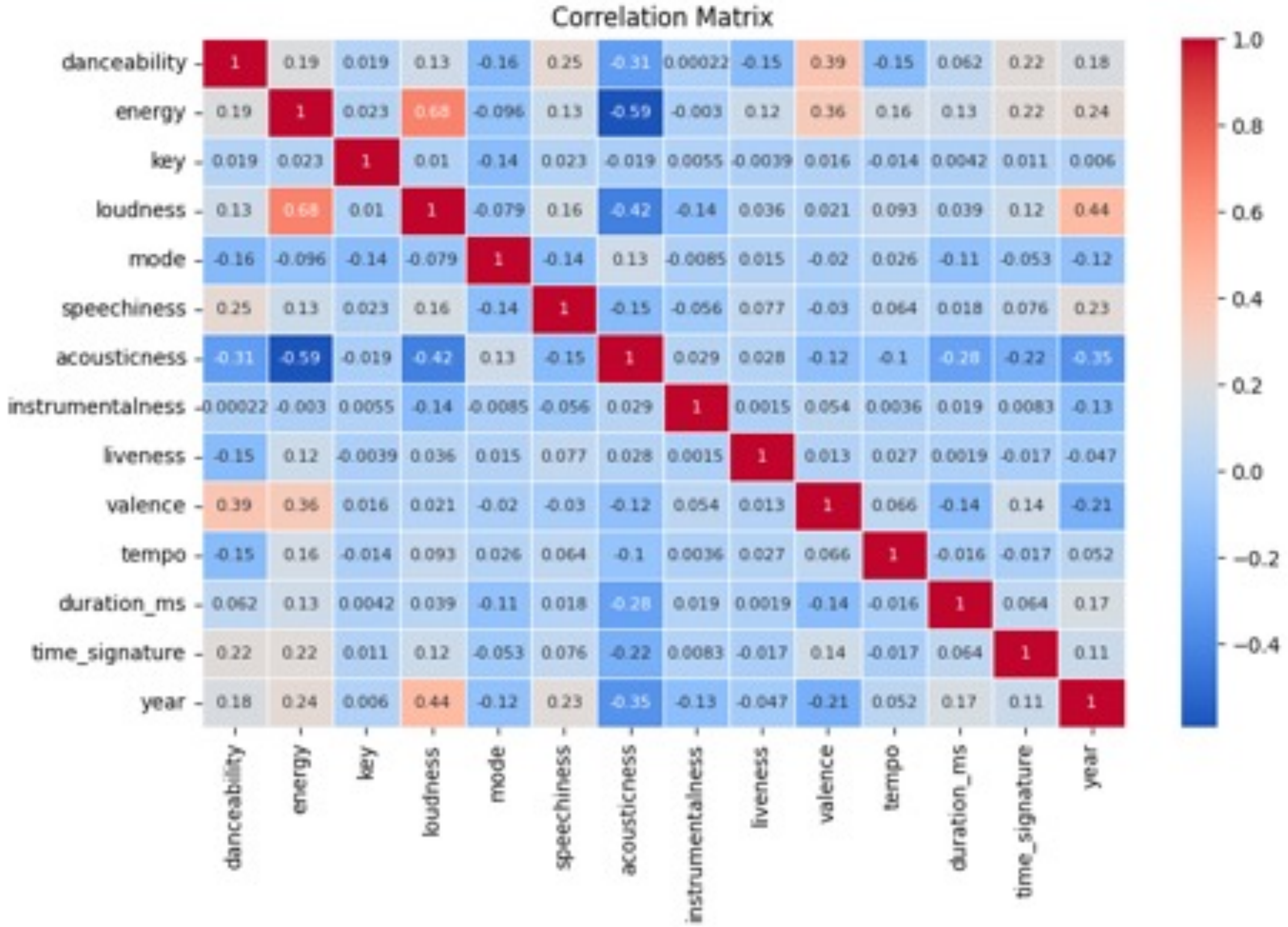



# FEATURE EXTRACTION AND EDA

Correlation of the Features:

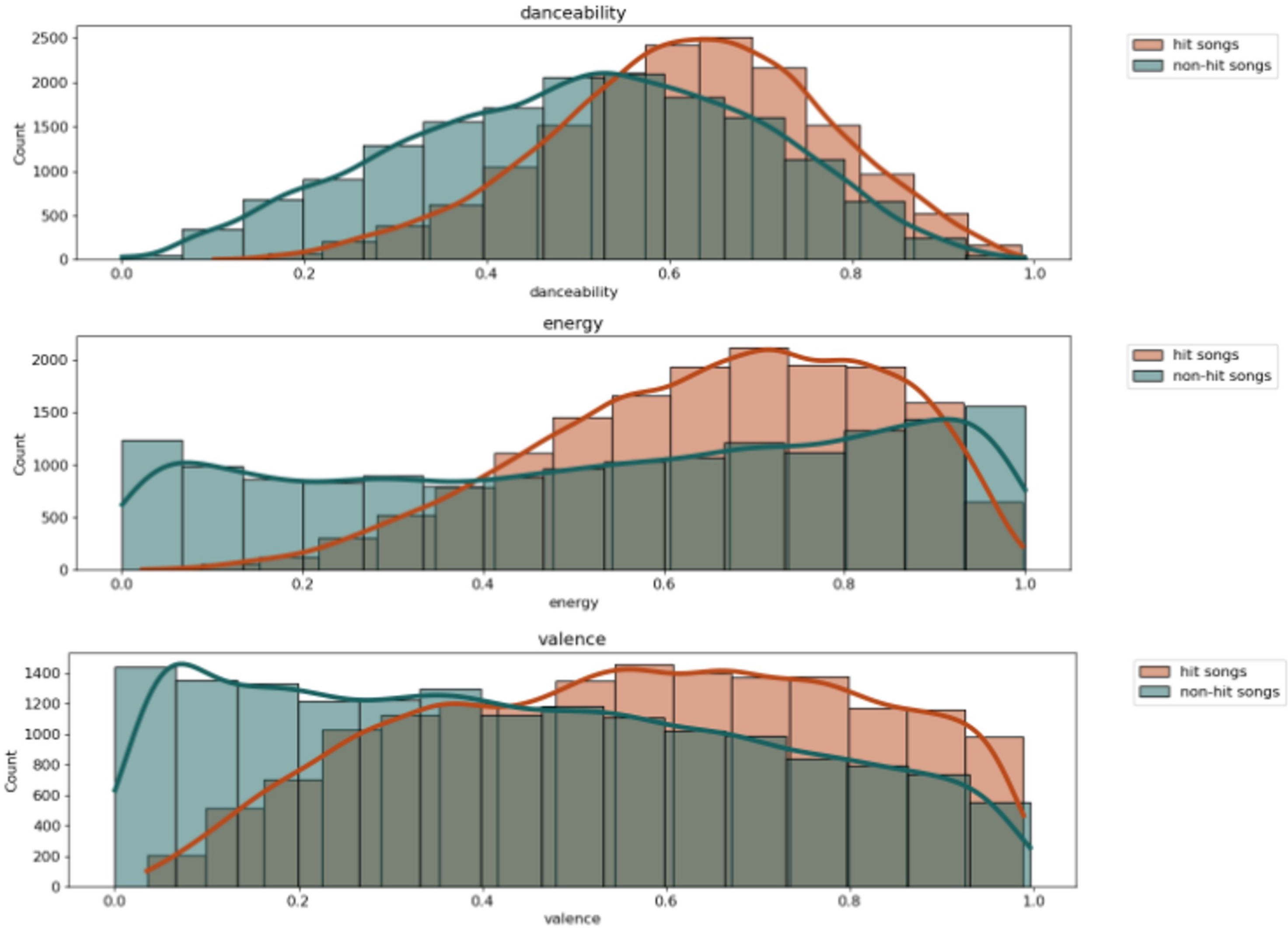
- **Energy** and **Loudness** are highly positively correlated
- **Acousticness** is negatively correlated with Energy and loudness



# FEATURE EXTRACTION AND EDA

High Priority Features:

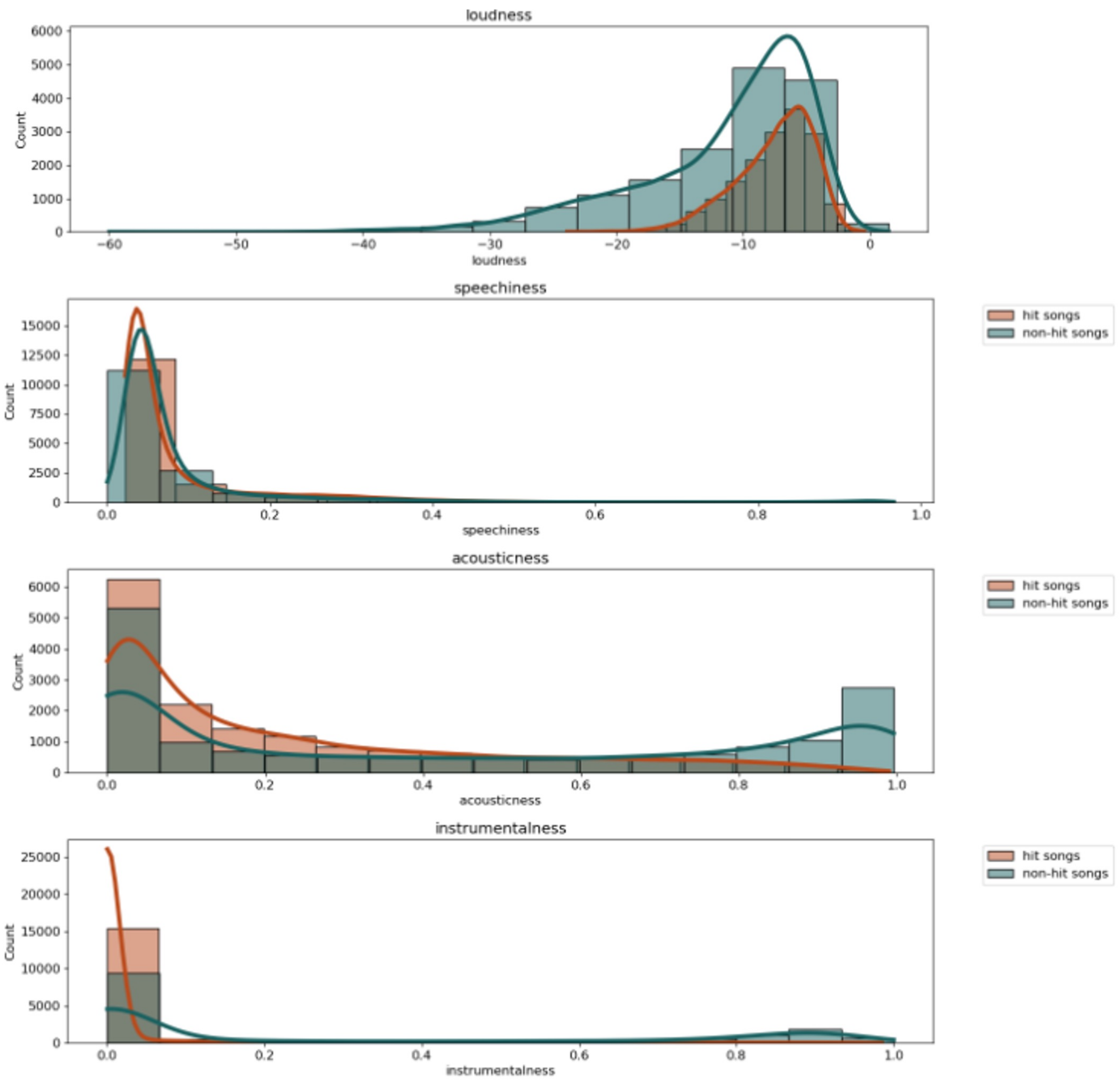
- **Danceability** - Top 100 slight peak shift to the right, differentiated trend from non-hits
- **Energy** - Top 100 larger peak shifted right while non-hits are more evenly distributed
- **Valence** - Top 100 and non-hits essentially have opposite distributions/peaks with hit songs trending to the right. There is a lot of overlap here though



# FEATURE EXTRACTION AND EDA (CONT.)

Mid-Priority Features:

- **Loudness** - Top 100 much tighter range compared to non-hits, not primary focus but could be good addition
- **Speechiness** - Top 100 is very tight lower range, non-hits also skew lower but have a bit more distribution, not primary focus but could be good add-on feature
- **Acousticness** - non-hits have strong peak towards upper range
- **Instrumentalness** - Top 100 all at low end, literally no variation. non-hits also favor low end but has an overall wider distribution of values



# DETAILS ON MODELS USED

## Logistic Regression (LR)

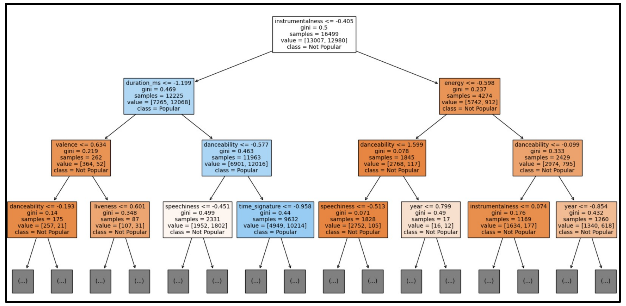
- LR model training:
  - *Used all the features in the dataset*
- Extracted the coefficients of each feature*
- Feature concentration check:
  - *Retrained the model using the 5 features with the highest absolute values of coefficients*

| | Feature | Coefficient | Absolute_Coefficient |
|---|---|---|---|
| 7 | instrumentalness | -1.189548 | 1.189548 |
| 3 | loudness | 0.763486 | 0.763486 |
| 1 | energy | -0.512679 | 0.512679 |
| 6 | acousticness | -0.427913 | 0.427913 |
| 0 | danceability | 0.381690 | 0.381690 |

* Top 5 features w/ the *highest absolute values of coefficients*

## Random Forest (RF)

- RF model training:
  - *Used all the features in the dataset*
- Single decision tree plot**
- Feature importance check
- Hyperparameter tuning:
  - *Randomized Search with Cross-Validation*

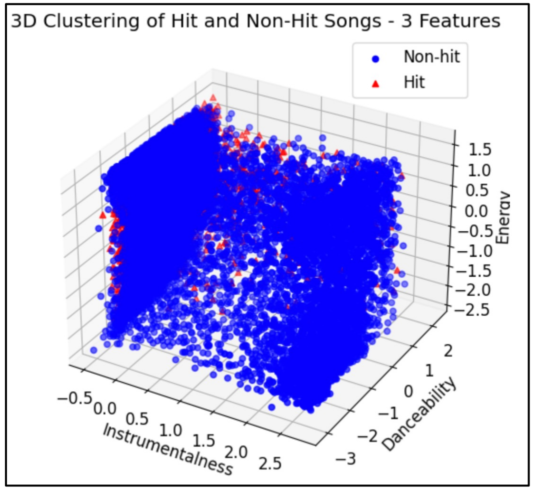



** One single decision tree

# DETAILS ON MODELS USED (CONT.)

## K-Nearest Neighbors (KNN)

- KNN model training:
  - *Used all the features in the dataset*
- 3D clustering of hit and non-hit songs*
  - *Selected three features w/ high importance and low collinearity*




* 3D clustering of hit and non-hit songs

## XGBoost

- XGBoost model training:
  - *Used all the features in the dataset*
- Hyperparameter tuning**:
  - *Randomized Search with Cross-Validation*
- Feature importance check

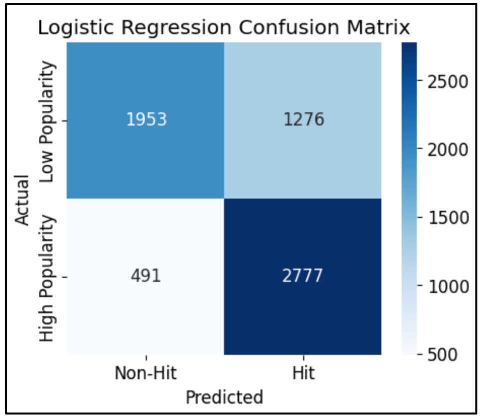
```
Fitting 3 folds for each of 8 candidates, totalling 24 fits
▸ RandomizedSearchCV
▸ estimator: XGBClassifier
    ▸ XGBClassifier
```

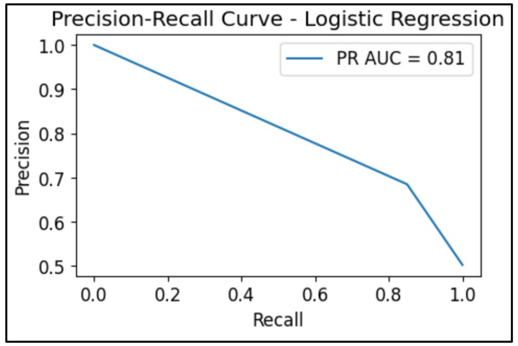
** Hyperparameter tuning using RandomizedSearchCV

# RESULTS AND OBSERVATIONS

## Logistic Regression (LR)

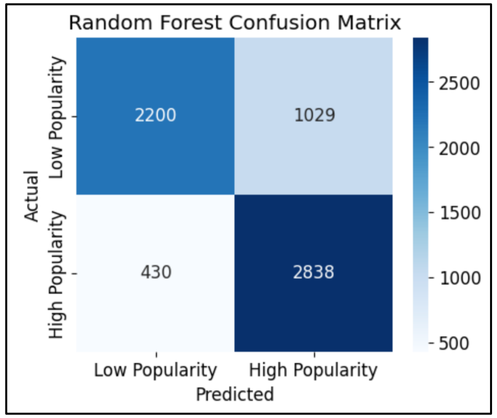
- Accuracy
  - *Initial model: ~ 72.74%*
  - *Retrained model: ~ 71.60%*
- Performance of initial model
  - *Confusion Matrix*
  - *Precision and Recall*

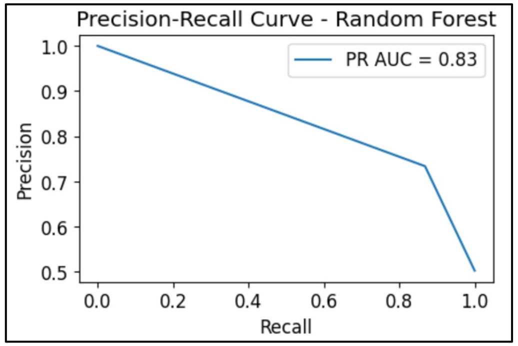



## Random Forest (RF)

- Accuracy
  - *Initial model: ~ 77.16%*
  - *Optimized model: ~ 77.25%*
- Performance of *optimized* model
  - *Confusion Matrix*
  - *Precision and Recall*

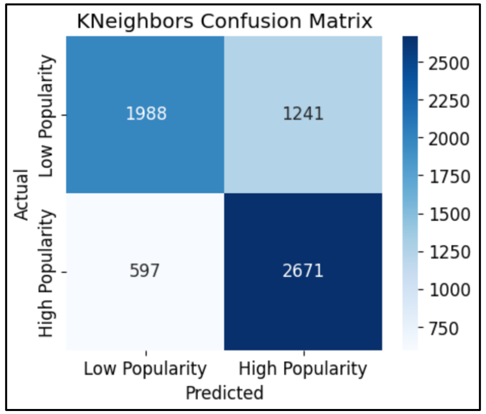


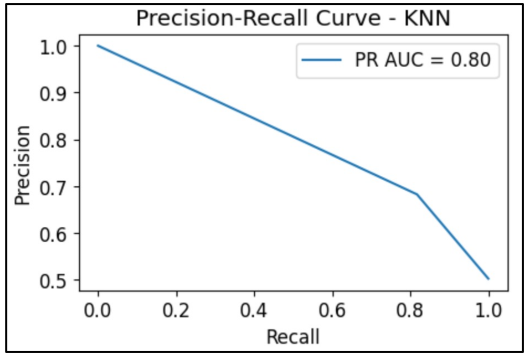



# RESULTS AND OBSERVATIONS (CONT.)

## K-Nearest Neighbors (KNN)

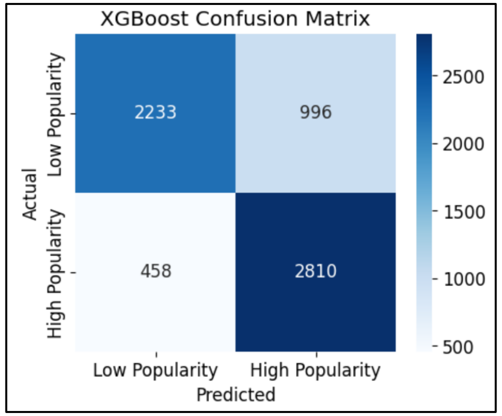
- Accuracy
  - *~ 72.52%*
- Performance
  - *Confusion Matrix*
  - *Precision and Recall*

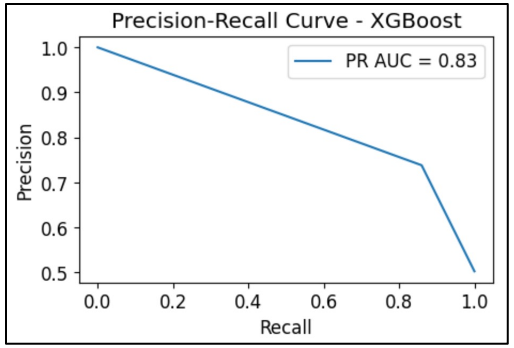



## XGBoost

- Accuracy
  - *Initial model: ~ 77.17%*
  - *Optimized model: ~ 78.08% (Highest)*
- Performance of *optimized* model
  - *Confusion Matrix*
  - *Precision and Recall*
  - *Learning Curve Plot (see slide 16)*

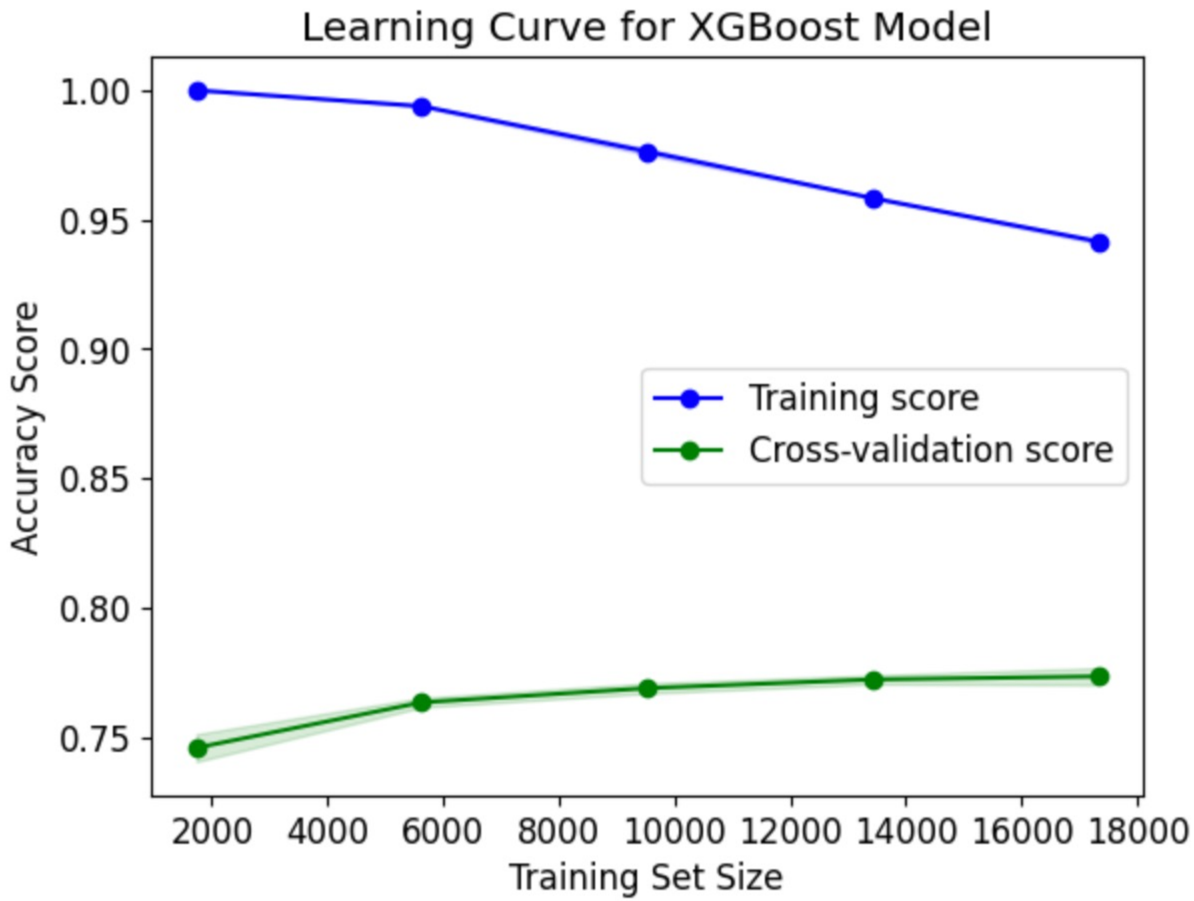


# RESULTS AND OBSERVATIONS (CONT.)

**Learning Curve of the Optimized XGBoost Model**

- Limitation:
  - *A complex model requires a larger amount of data to start generalizing well.*
  - *The training score and the cross-validation score have a trend to converge if more training data is available, which indicates the model may generalize better with a larger dataset.*



# NEXT STEPS AND RISKS

Try other models

- Neural Networks
- Unsupervised clustering for feature engineering

Risks:

- Model overfitting
- Computational efficiency

# REFERENCES


1. Amitansh Joshi, Amit Parolkar, Vedant Das. (2023). Spotify_1Million_Tracks [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/59878
2. T. Li, M. Choi, K. Fu and L. Lin, "Music Sequence Prediction with Mixture Hidden Markov Models," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 6128-6132, doi: 10.1109/BigData47090.2019.9005695.
3. Dhruvil Dave. (2021). Billboard "The Hot 100" Songs [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/1211465
4. Dimolitsas, Ioannis & Kantarelis, Spyridon & Fouka, Afroditi. (2023). SpotHitPy: A Study For ML-Based Song Hit Prediction Using Spotify.
5. Figueroa Rodolfo. (2023). Spotify 1.2M+ Songs [Data set]. Kaggle. Spotify 1.2M+ Songs (kaggle.com)
6. D. Herremans, Martens, D., and Sörensen, K., “Dance hit song prediction”, Journal of New music Research, vol. 43, no. 3, p. 302, 2014.
7. Ioannis Karydis, Aggelos Gkiokas, Vassilis Katsouros. Musical Track Popularity Mining Dataset. 12th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), Sep 2016, Thessaloniki, Greece. pp.562-572, ⟨10.1007/978-3-319-44944-9_50⟩. ⟨hal-01557622⟩
8. Middlebrook, Kai & Sheik, Kian. (2019). Song Hit Prediction: Predicting Billboard Hits Using Spotify Data.
9. Miller Sean. (2021) Billboard Hot weekly charts [Data set]. Data World, Kaggle. Billboard Hot weekly charts - dataset by kcmillersean | data.world
10. Ni, Y., Santos-Rodriguez, R., Mcvicar, R., De Bie, T.: Hit song science once again a science. In: 4th International Workshop on Machine Learning and Music (2011)
11. Pachet, F. (2011) Hit Song Science. In Tao, Tzanetakis & Ogihara, editor, Music Data Mining, CRC Press/Chapman Hall
12. Pachet, F. and Roy, P. (2008) Hit Song Science is Not Yet a Science. Proceedings of Ismir 2008, pages 355-360, Philadelphia, USA
13. A. H. Raza and K. Nanath, "Predicting a Hit Song with Machine Learning: Is there an apriori secret formula?," 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia, 2020, pp. 111-116, doi: 10.1109/DATABIA50434.2020.9190613.
14. Spotify. Spotify for Developers. Retrieved January 19, 2024, from https://developer.spotify.com/documentation/
15. L. -C. Yang, S. -Y. Chou, J. -Y. Liu, Y. -H. Yang and Y. -A. Chen, "Revisiting the problem of audio-based hit song prediction using convolutional neural networks," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 621-625, doi: 10.1109/ICASSP.2017.7952230.
16. Zhao, M., Harvey, M., Cameron, D., Hopfgartner, F., Gillet, V.J. (2023). An Analysis of Classification Approaches for Hit Song Prediction Using Engineered Metadata Features with Lyrics and Audio Features. In: Sserwanga, I., et al. Information for a Better World: Normality, Virtuality, Physicality, Inclusivity. iConference 2023. Lecture Notes in Computer Science, vol 13971. Springer, Cham. https://doi.org/10.1007/978-3-031-28035-1_21